## 

R Practice

ALY6040 - Data Mining

Northeastern University

Chongfan Bao

Professor: Justin Grosz

May 30, 2021

**Abstract**

A successful business needs to dig out useful data to help an enterprise or leader to make decisions and promote the development of the enterprise. The dataset is related to sell ice cream. The researcher leverages the RStudio to explore the latent information in order to know the habits of customers. Lastly, based on data visualization, the entrepreneur can know how to make a selling strategy.

*Keywords:* data exploration, data visualization, business strategy

**Introduction**

In this module, the researcher will explore the dataset consisted of 10000 records and 11 variables, which embraces the string and numeric type. The data depicts the situation of donation for Kick Starer, such as the date, sex, household income, and preferred color of the device. In addition, the data also shows the number of ice cream that people eat per week, the flavor of ice cream. However, before conducting data analysis, the researcher notices that the part of the data is missing or has outliers. To clean data is one of the most important aspects of data science. As a data scientist, you can expect to spend up to 80% of your time cleaning data (John, 2019). During this process, the researcher will use various packages to date clean data and visualize data so as to find latent business information and assist make strategies.

**Data Analysis**

Figure1 depicts a new data structure that people notice that the raw data includes 10000 records and 11 variables. The date format is presented as Tibble in RStudio, which is similar to the data frame. The variables incorporate three formats: number, character, and date and time.

**Text

Description automatically generated with medium confidence**

**Figure 1:** *The structure of data*

From figure2 the researcher can know that some data are missing and the missing numbers of data. For instance, the variable named “Deposit Amount” has 40 missing values. Moreover, people can know the maximize, minimize, mean, median values of numerical variables that is beneficial to clean data.

*Table

Description automatically generated*

**Figure 2:** *Summary data structure*

In figure 3, the researcher uses the box& whisker plot to detect abnormal values. In addition, the researcher also leverages the is.na and filter functions to omit outliers. The mutate function is leveraged to replace NA values with mean and median.

*Graphical user interface

Description automatically generated with medium confidence*

**Figure 3:** *Detecting* *anomalies*

Figure 4 and figure 5 displays the unpolluted data after cleaning. There are not any NA values and anomalies.

Diagram

Description automatically generated

**Figure 4:** *After cleaning data*

*Graphical user interface, text, application, email

Description automatically generated*

**Figure 5:** *Detecting outliers**after cleaning data*

From figure 6, people can know that women prefer to buy ice cream the male. Usually, people will consume at least 5 ice cream products.

*Chart, bar chart

Description automatically generated*

**Figure 6:** *The**histogram of ice cream products consumed per week*

Based on figure 7, people apparently don’t have extreme preferences about the color of the device.

*Chart, bar chart

Description automatically generated*

**Figure 7:** *The preferred color of the device*

According to chart 8, the female likes swirl ice cream. the vanilla ice cream is popular for males and females. Therefore, the merchandiser can make more swirl and vanilla ice cream.

**Chart, bar chart

Description automatically generated**

**Figure 8:** *The preferred color of the device*

According to chart 9, people notice that there are not any male consumer records, which also means that most males only eat no more than 3 ice cream a week. The female is mainly consumers.

**Chart, bar chart

Description automatically generated**

**Figure 9:** *The number of desserts is eaten a week*

According to chart 10, The households whose incomes are more than 100k are willing to donate money to Kick Starter. Hence, people can target these people to raise money and it will be easier to get money.

**Chart, bar chart, box and whisker chart

Description automatically generated**

**Figure 10:** *The donation situation based on household income*

**Conclusion**

After approximately data exploration, the female prefers to buy the ice cream, especially vanilla and swirl ice cream. Most males won't eat ice cream more than 3 a week. So the vendors can place stores or sell ice cream adjacent to places that women are willing to go, such as shopping malls, boutiques, or theaters. The suppliers should remain at least 7 ice cream machines in the warehouse. This is because there are 5 ice cream machines sold per week. The different color machines have at least one in the warehouse. Lastly, the aim to get donations should be households whose incomes are more than 100k.

**Next Steps**

In the future work, the researcher will explore which month of the year is better for raining money, and what gender prefers to donate money. In addition, the researcher is going to find out what kinds of income households want to buy a Keurig. All of these researches contribute to make an effective marketing strategy

**References**

Hadley, W., & Garret, G., (2017). R for Data Science. Retrieved from https://r4ds.had.co. nz/index.html

John, S., (2019). Data Cleaning with R and the Tidyverse: Detecting Missing Values. Retrieved from https://towardsdatascience.com/data-cleaning-with-r-and-the-tidyverse-detecting-missing-values-ea23c519bc62

Karlijn, W., (2017). Pipes in R Tutorial For Beginners. Retrieved from https://www.datacamp. com/community/tutorials/pipe-r-tutorial

**Appendix**

*The R code*

*Graphical user interface, text, application

Description automatically generatedGraphical user interface

Description automatically generated*